Università di Genova
DIPARTIMENTO
INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

Computer Science Workshop
PhD program in Computer Science and Systems Engineering

Looking At Each Other (LAEO)

Federico Figari Tomenotti, Giorgio Cantarini, Nicoletta Noceti, Francesca Odone

Preliminary work towards people interaction understanding

Introduction

In this work, we focus on one of the main visual cue at our disposal for human interaction understanding. It is of uttermost relevance to consider as an important event the moment when two people look at each other. Every types of interaction starts (and hopefully ends) with a mutual gaze, it is also a proxy for taking turn in actions and particularly in speaking [2].
To infer the direction of view of a person we introduce a novel method to estimate the head pose of people starting from a small set of head keypoints. To this purpose, we propose a regression model that exploits keypoints computed automatically by 2D pose estimation algorithms for static images and outputs the head pose represented by yaw, pitch, and roll.
Afterwards, this piece of information id fed to an algorithm who calculate eye-interaction between people and tracks how much time this interaction persists and how strong it is.

Main Objectives

1. Head pose estimation as proxy for gaze direction.
2. Detection of LAEO event.
3. Tracking of LAEO event duration and people involved.
4. Usage of the head pose *uncertainty* for an accuracy refinement for the detection event.

Materials and Methods

Head Pose Estimator The model proposed is based on a Bayesian neural network trained on different datasets (BIWI, AFLW-2000, 300W-LP), both real and synthetic. It is simple to implement and more efficient with respect to the state of the art – faster in inference and smaller in terms of memory occupancy – with comparable accuracy. But moreover, it provides a confidence measure on its predictions learned through an appropriately designed loss function. The model takes in input five keypoints (extracted via a Human-Pose estimation algorithm, such as Openpose or Centernet) of the head and output three angle (yaw, pitch and roll) and one uncertainty measure for each.
LAEO detector The second part of the algorithm, uses this piece of information, and the results from the Human-Pose estimator to calculate via geometric interpretation a measure of how much a person is looking towards another one, and then combine all the measure in a frame to state if a couple (or more) of people are looking at each other. This type of measure can also be extended over time, using also a tracking algorithm and some type of filtering in the time dimension. We tested our solution comparing against a dataset provided in [4] for the discoveries of LAEO pairs in videos.

LAEO Algorithm

```
$\mathbf{u}_{AB} \leftarrow (x_B - x_A, y_B - y_A)$
$\mathbf{h}_A \leftarrow (x'_A - x_A, y'_A - y_A)$
$\mathbf{h}_B \leftarrow (x'_B - x_B, y'_B - y_B)$
$cos(\alpha_A) \leftarrow \frac{\mathbf{u}_{AB} \cdot \mathbf{h}_A}{|\mathbf{u}_{AB}| \cdot |\mathbf{h}_A|}$
$cos(\alpha_B) \leftarrow \frac{-\mathbf{u}_{AB} \cdot \mathbf{h}_B}{|\mathbf{u}_{AB}| \cdot |\mathbf{h}_B|}$
$LAEO_{value} = w_A cos(\alpha_A) + w_B cos(\alpha_B)$
Return $LAEO_{value}$
```

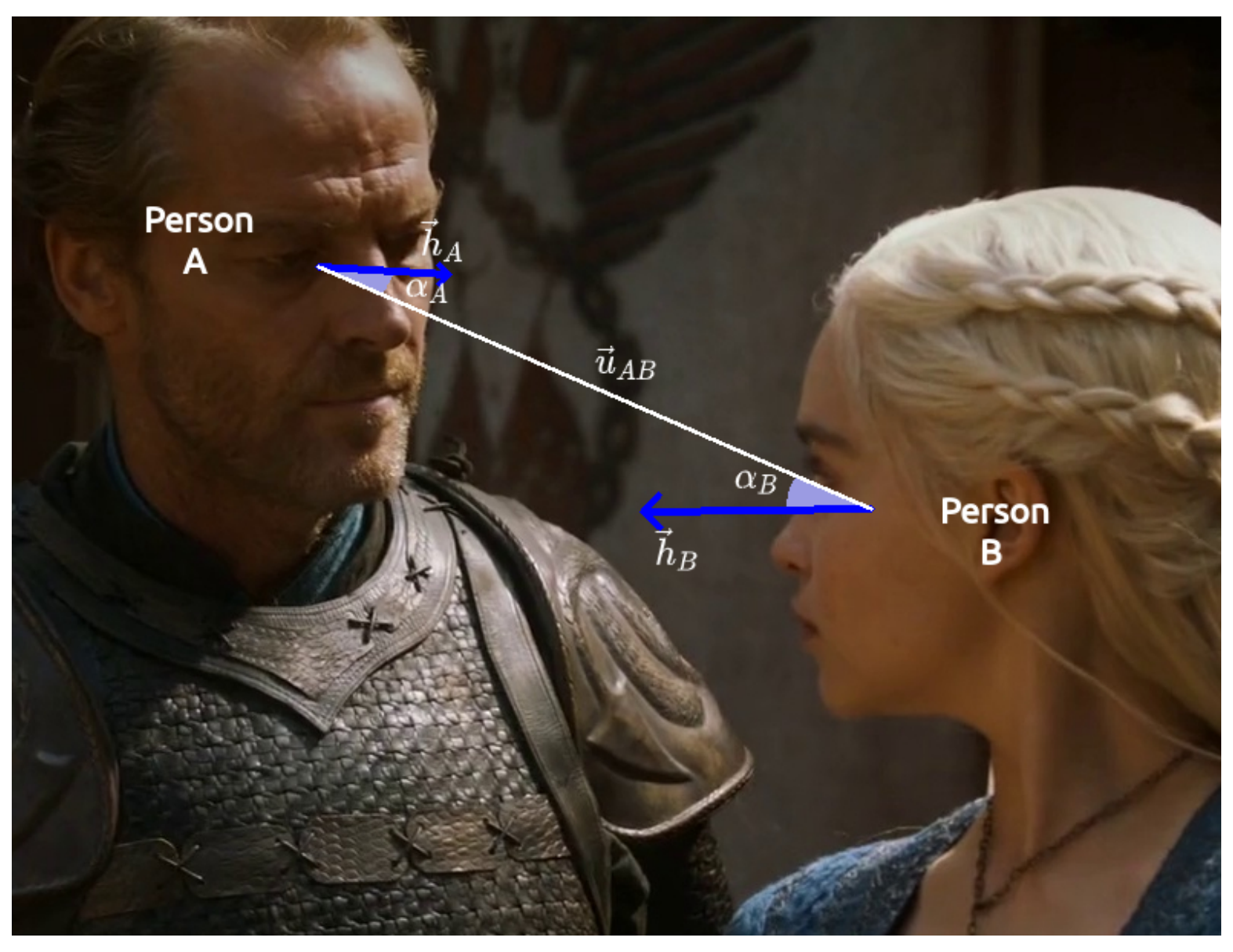

Figure 1: Example and explanation of variables in the algorithm.

Experimental Section

Even if our LAEO algorithm is more complex, we evaluated it only on a dataset proposed in [4], which is composed of 100 videos from movies, and each frame is annotated with a couple of LAEO people. As long as our algorithm is able to find more than one LAEO couple per frame, we did adopt more evaluation routines. The first one is simply to check if in our detections the ground truth couple is present; the second checks if the ground truth couple has been found to have the strongest interaction, and the last checks if we labelled a frame as LAEO where it is stated by the ground truth, disregarding which couple has been retrieved.

Here are reported the results for the last and easy experiment combined with the first: retrieving a LAEO interaction in a frame, it was considered valid if and only if the same couple were present in the ground truth.

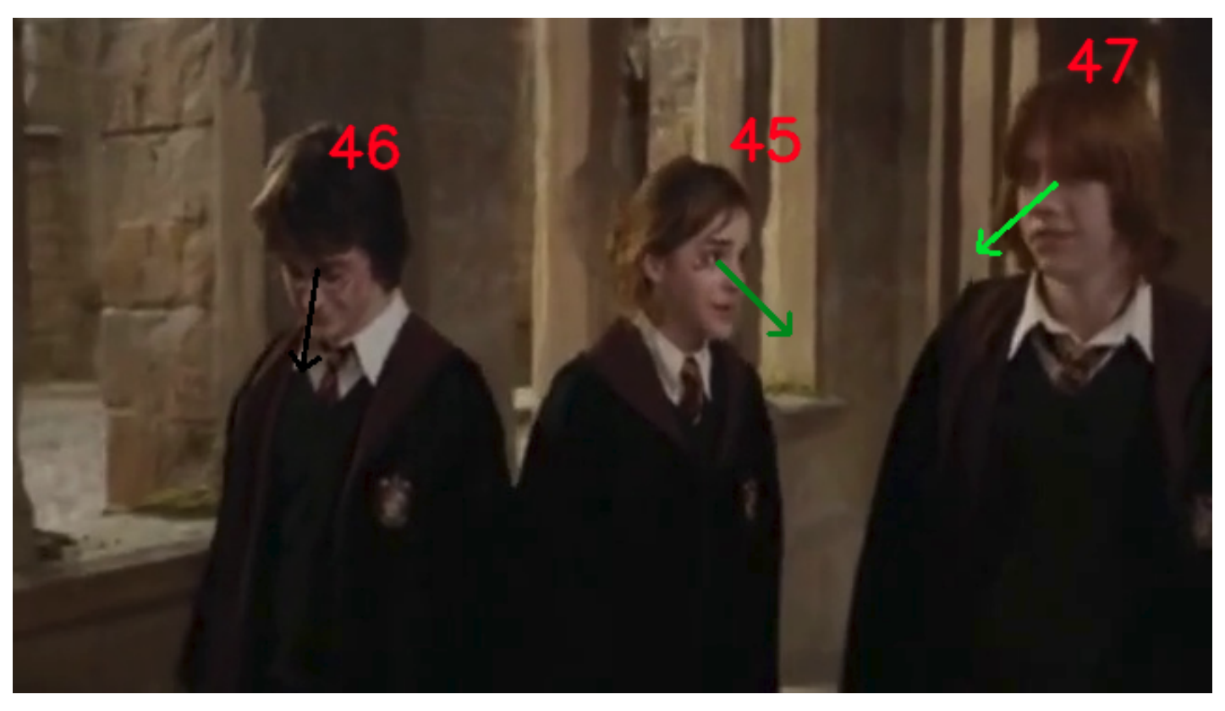

Figure 2: Example of frame with retrieved arrows representing gaze directions. In black arrows looking at no one. In green arrows looking towards another person (the more intense the colour, the more the interaction) In red id for each person.

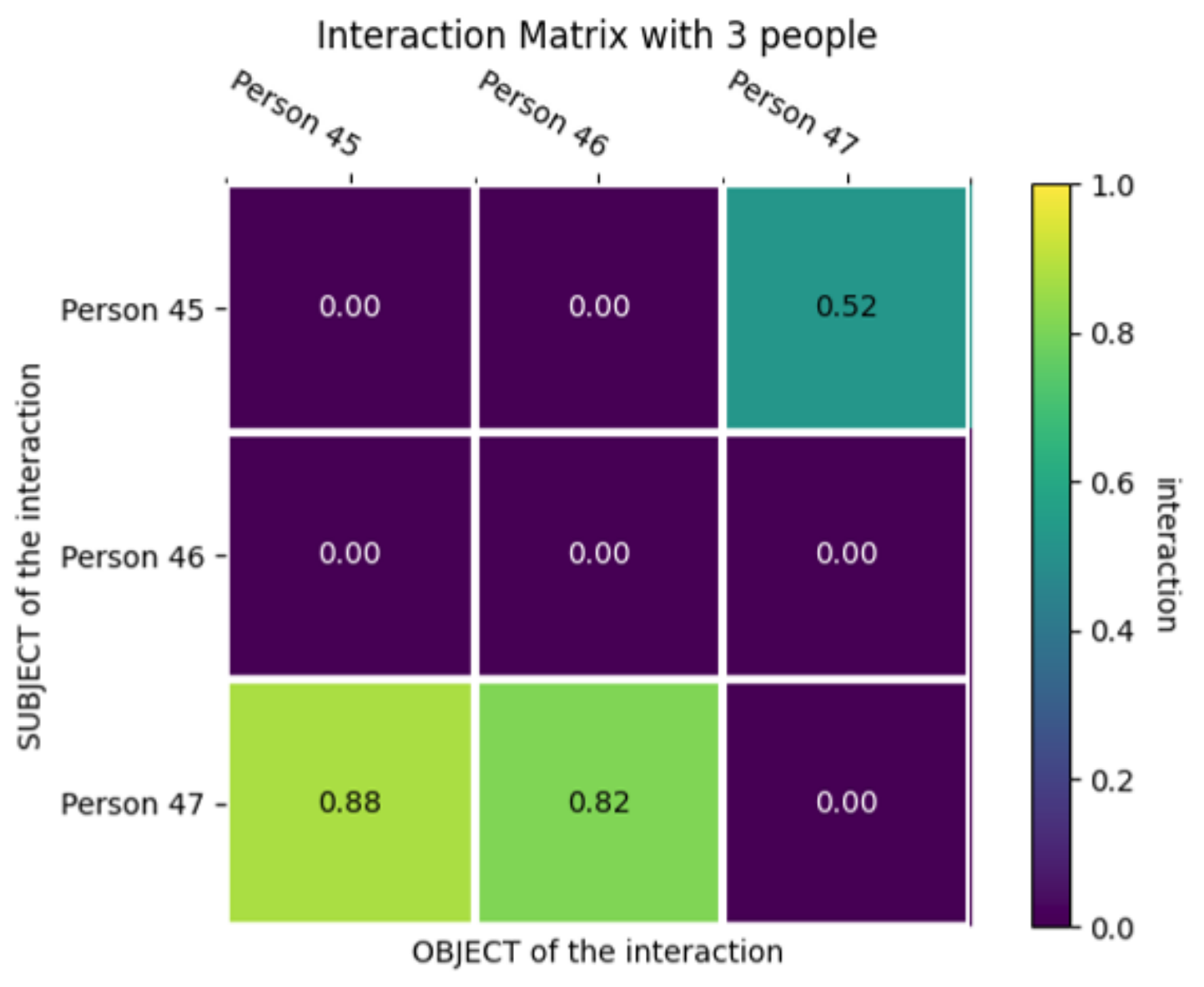

Figure 3: Matrix representing interactions for the figure beside.

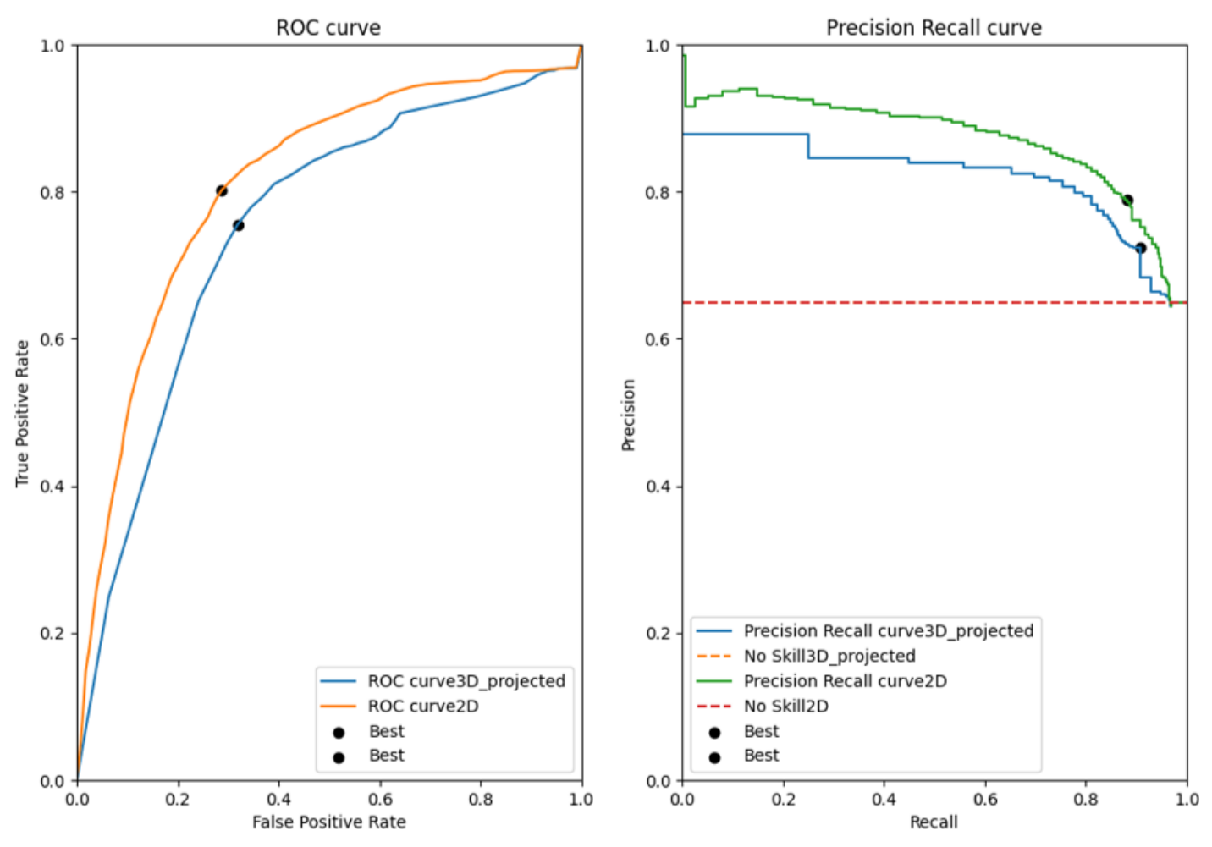

Figure 4: Left: ROC curve for 3D-projected angle (the current work) and 2D angle [1]. Right: P-R curve.

Method	**PREC**	**REC**	**F**	**AP**
LAEO-Net [4]	–	–	–	0.80
LAEO-Net++ [3]	–	–	–	0.87
Ours Baseline	0.77	0.80	0.78	0.86
Ours with uncert.	0.80	0.72	0.76	0.88

Table 1: The performance of our method for LAEO detection on the UCO-LAEO dataset. AP is estimated as in [3].

Conclusions

The present work is ongoing, but results show up, performing in the simple LAEO task as one deep learning algorithm trained specifically for the purpose. The overall accuracy, both in terms of single LAEO instances and on LAEO people retrieval (i.e. who are the peope LAEO in the frame?) are comparable with the state of the art. To conclude, a real time version of the algorithm has been developed and proved working.

Forthcoming Research

- move towards a 3D algorithm to exploit the Head Pose Estimator, use of 3D data
- use the information as prior in a robust network for LAEO classification
- expand towards general human interaction understanding for small groups
- study the usage of the uncertainty information

References

[1] P. A. Dias, D. Malafronte, H. Medeiros, and F. Odone. Gaze estimation for assisted living environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 290–299, 2020.

[2] S. Ho, T. Foulsham, and A. Kingstone. Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PloS one*, 10(8):e0136905, 2015.

[3] M. Marín-Jiménez, V. Kalogeiton, P. Medina-Suárez, and A. Zisserman. Laeo-net++: revisiting people looking at each other in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2020.

[4] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019.

The research has been carried on thanks an AFOSR grant project in the MaLGa Laboratory.

CONTACTS

Federico Figari Tomenotti
federico.figaritomenotti@edu.unige.it